Optical Flow Techniques for Facial Expression Analysis: Performance Evaluation and Improvements

Benjamin Allaert, Isaac Ronald Ward, Ioan Marius Bilasco, Chaabane Djeraba, Mohammed Bennamoun

Optical Flow Techniques for Facial Expression Analysis: Performance Evaluation and Improvements

Benjamin Allaert [1] *, Isaac Ronald Ward [2], Ioan Marius Bilasco [1], Chaabane Djeraba [1] and Mohammed Bennamoun [2]

[1] Centre de Recherche en Informatique Signal et Automatique de Lille, Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL -, F-59000 Lille, France.
[2] Department of Computer Science and Software Engineering, The University of Western Australia, Perth, Australia.
* Corresponding author: benjamin.allaert@univ-lille.fr, ORCID: 0000-0002-4291-9803.

Abstract: Optical flow techniques are becoming increasingly performant and robust when estimating motion in a scene, but their performance has yet to be proven in the area of facial expression recognition. In this work, a variety of optical flow approaches are evaluated across multiple facial expression datasets, so as to provide a consistent performance evaluation. Additionally, the strengths of multiple optical flow approaches are combined in a novel data augmentation scheme. Under this scheme, increases in average accuracy of up to 6% (depending on the choice of optical flow approaches and dataset) have been achieved.

Keywords: optical flow, facial expression, deep learning, data augmentation.

---

1. Introduction

Building a system that is capable of automatically recognizing the emotional state of a person from their facial expressions has been a burgeoning topic in computer vision in the recent years. Automating the analysis of facial expressions, from videos, is highly beneficial in a range of diverse applications, including security, medicine, and human-machine interaction. For instance, the analysis of the emotional state of a patient, based on their facial expressions, can help to estimate the quality of the provided care, and to monitor the ongoing patient-doctor relationship.

The use of facial expression information increases proportionally with our need to automate the process of extracting behavior and cognitive-related information (expressions, intentions, predictions). Although many advances have been achieved in this area, the recent approaches do not yet achieve satisfactory results when deployed in real-world situations (e.g., in transportation and retail stores). Indeed, the data acquired in unconstrained settings is highly heterogeneous. Complications encountered in unconstrained settings highlight several scientific problems which remain unresolved; illumination changes, variations in pose, facial occlusions and variations in expression intensity which may considerably affect a system's performance.

The analysis of the facial movement through optical flow seems to offer a promising avenue of research for expression analysis. It is mainly used for its ability to characterize both intense and subtle movements [1], as well as being able to correct head pose variations [2], or to deal with partial facial occlusions [3].

A number of methodological innovations have progressively been introduced to improve the performance of optical flow techniques on datasets, such as MPI-Sintel [4], as illustrated in the top of Figure 1. However, several authors suggest that the use of the *recent* optical flow approaches tends to *reduce* the system performance in fields such as human action recognition [5] or facial expression recognition [6]. That is, in comparison with the more common optical flow techniques — such as the one proposed by Farnebäck [7] (reflected in the bottom of Figure 1). Nevertheless, no clear protocol of comparison has yet been proposed.

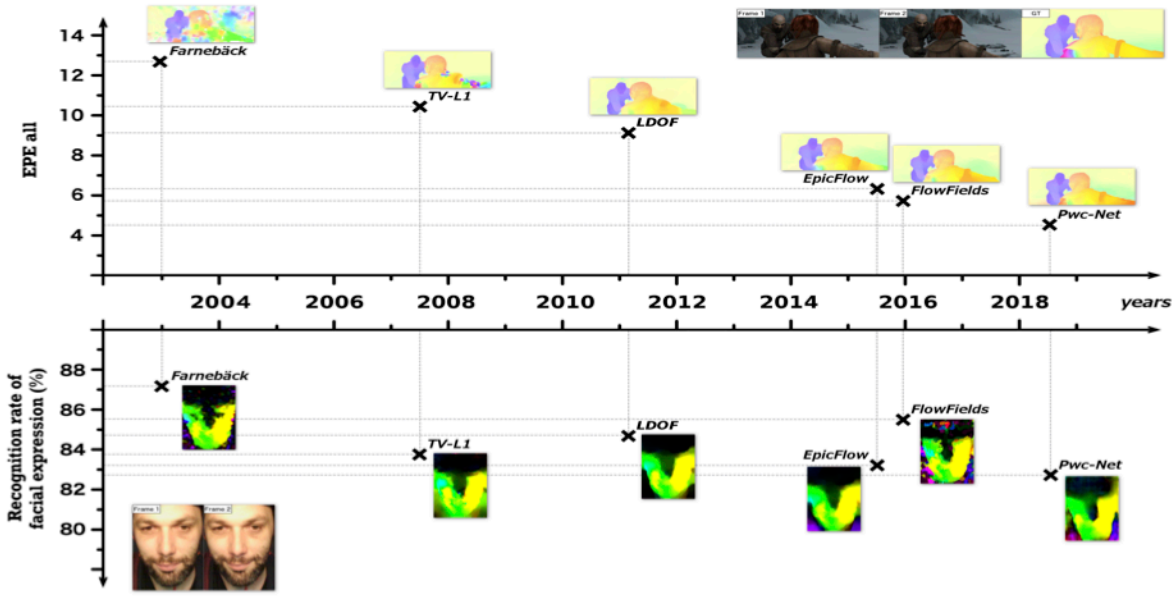

Figure 1. Comparison of the performance of several optical flow approaches on the MPI-Sintel (top) and on a set of facial expression datasets (bottom). Such datasets are used due to the lack of optical flow ground truth data for facial expression analysis. Although the performance tends to be conclusive on MPI-Sintel, this is not the case for facial expression analysis, where a basic approach such as Farnebäck gives the best performance.

To understand this paradox, this work investigates the impact of the most recent dense optical flow approaches on the performance of facial expression recognition. More specifically, this study is the first attempt to address the question: "Which optical flow approach should one use to analyze facial motion?".

The paper is organised as follows: we introduce, in Section 2, the challenges of detecting facial motion (especially motion discontinuities), and we briefly describe the main characteristics of the major optical flow techniques that are proposed in the literature. In Section 3, we introduce the datasets that we used to compare our selected optical flow approaches and define their performance criteria. We then evaluate the capacity of our selected optical flow approaches to accurately detect facial movements, by combining them with different hand-crafted and learning-based approaches on a variety of other facial expression datasets in Section 4. In Section 5, we analyze the use of distinctive features of different optical flow approaches to artificially increase the learning data. To conclude, we summarize our results and discuss future perspectives in Section 6.

2. Scope and background

This section highlights the main objectives of the paper and gives a brief overview of some current optical flow approaches and their characteristics.

2.1. Scope of the paper

The unique characteristics of facial movement implies that some motion discontinuities tend to provide information about an expression [8]. Therefore, the need to devise dense optical flow approaches to address motion discontinuities, while ensuring a rapid computation time, is both an important requirement and a significant challenge. Consequently, it is important to study how optical flow techniques deal with motion discontinuities, while being immune to noise propagation in the neighboring regions.

Although optical flow approaches are becoming more and more robust on datasets such as MPI-Sintel, it is important to consider the performance of these approaches for facial expression analysis. As stated, the challenges proposed by the data of MPI-Sintel do not always reflect the problems that can be observed on a face. In this context, relying on the results obtained by optical flow approaches on MPI-Sintel may not be relevant in identifying the best optical flow approaches to characterize facial motion.

In this paper we provide three key contributions. **First**, we study the ability of different optical flow approaches in characterizing facial expressions. The key processing steps of each approach are analyzed in order to identify those which have a tendency to improve or reduce performance. **Second**, we investigate whether several optical flow approaches can be used to characterize facial movements, in place of using a single, highly performant approach. Put more explicitly, we answer the following question: "Can optical flow approaches be used in a data augmentation process in the context of deep learning architectures?" To answer this, we analyze several optical flow approaches and their characteristics — taking particular note of how they handle motion discontinuities. **Finally**, in order to benchmark and compare the results of our work, we propose a new evaluation baseline for evaluating the performances obtained by using *several* optical flows on various facial expression datasets, by comparing different hand-crafted and learning-based algorithms.

In order to ensure that our experiments are reproducible, all the data are made available online .*

2.2. Background of the optical flow techniques

Optical flows are relatively sensitive in the presence of some factors such as occlusions, light changes or out-of-plane movement. All these factors lead to the appearance of false movements, which result in motion discontinuities. Training benchmarks such as MPI-Sintel [4] have been proposed in order to address these problems. In response, many optical flow approaches have been proposed. Some of these approaches are

*Download link for reproducibility : [...]

distinguished by their originality in regards to how they implement some key processing steps, including the matching, filtering, interpolation and optimization.

Most optical flows typically initialize using sparse descriptor matching techniques or by dense approximate fields in the nearest neighboring fields [7,9,10]. The techniques which are of interest to this work are described *briefly* here. For a quantitative comparison, please see Section 4.

Farnebäck's method [7] embeds a translation motion model between neighborhoods of two consecutive images in a pyramidal decomposition. Polynomial expansion is employed to approximate pixel intensities in the neighborhood. The tracking begins in the lowest resolution level, and continues until convergence. The pyramid decomposition enables the algorithm to handle large pixel motions, including distances greater than the neighborhood size. Clearly this algorithm cannot compete with the more recent methods which highly reduce discontinuities — however, as it does not include a generic post-processing step, it presents a good compromise between speed and accuracy.

Ldof [11] estimates large movements in small structures by integrating the correspondences (from descriptor matching) into a variational approach. These correspondences are *not* used in order to improve the accuracy of the approach; they are used as they support the coarse-to-fine warping strategy and avoid local minima.

TV-L1 [9] is a particularly appealing formulation which is based on total variation (TV) regularization and the robust L1 norm in the data fidelity term. This formulation can preserve discontinuities in the flow field and thus offers an increased robustness against illumination changes, occlusions and noise.

EpicFlow [12] relies on the deep matching algorithm integrated into the **DeepFlow** method [13] and interpolates a set of sparse matches in a dense manner to initiate the estimation. This approach preserves the edges so that they can be used in the interpolation of movement. The solution has proven its effectiveness in characterizing optical flows over multiple datasets, including MPI-Sintel [4] and KITTI [14].

Approaches which initialize via the use of dense approximate fields in the nearest neighbouring fields naturally have the advantage of being dense, but they have the major drawback of being highly outlier prone. This includes being prone to motion discontinuities.

Recent optical flow techniques such as **FlowFields** [15] use a dense correspondence field technique that is *much less* outlier prone. This method does not require explicit regularization or smoothing (such as in median filtering), but is instead a pure data-oriented search strategy which only finds the most inliers, while effectively avoiding the outliers.

PWC-net [16] is unlike recent learning-based approaches, and is based on a compact CNN model which uses simple and well-established principles: pyramidal processing, warping, and the use of a cost volume. The particularity of this method is that the warping and the cost volume layers have no learnable parameters that can reduce the model size. As with most recent approaches, motion discontinuities are handled by post processing the optical flow using median filtering.

3. Datasets and performances criteria

In this work, two primary sets of experiments are conducted: the evaluation of optical flow approaches (Section 4), and the augmentation of training data through use of optical flows (Section 5). The datasets used and the performance criteria in each case is outlined in this section.

3.1. Datasets

There is no dataset which offers a ground truth for accurately comparing the performance of optical flow approaches against the task of characterizing facial movements. Thus, we first propose a baseline based on a set of facial expression datasets which contain different expression intensities. For the purpose of our work, it is necessary to analyze temporal sequences, and hence image datasets such as JAFFE, RaFD or AffectNet are not considered. As the main aim of this study is to evaluate the capacity of the optical flow approaches in characterizing facial movement, we select data acquired in controlled conditions, where only the movement related specifically to the facial expression is present. Datasets such as MMI, DISFA, RECOLA and so forth, which contain numerous pose variations, occlusions and light changes are thus omitted from this study, as the biases induced by these challenges interfere with the native capacity of the optical flow approaches in characterizing facial movement.

We hence combine several datasets, specifically the CK+ [17], Oulu-CASIA [18] and SNaP-2DFe [19] datasets, which contain the six basic expressions (anger, disgust, fear, happiness, sadness, and surprise). A brief overview of each dataset is provided here for completeness:

CK+ contains 593 acted facial expression sequences from 123 participants, with seven basic expressions (anger, contempt, disgust, fear, happiness, sadness, and surprise). In this dataset, the expression sequences start in the neutral state and finish at the apex state. As illustrated in Figure 2, expression recognition is completed in excellent conditions, because the deformations induced by the ambient noise, facial alignment and intra-face occlusions are not significant with regard to the deformations that are directly related to the expression. However, the temporal activation pattern is variable in this dataset, and spreads from 4 images to 66 images with a mean sequence length of 17.8 ± 7.42 images.

Oulu-CASIA includes 480 sequences of 80 subjects taken under three different lighting conditions: strong, weak and dark illuminations. They are labeled with one of the six basic emotion labels (anger, disgust, fear, happiness, sadness, and surprise). Each sequence begins in the neutral facial expression state and ends in the apex state. Expressions are simultaneously captured in visible light and near infrared.

SNap-2DFe contains 1260 sequences of 15 subjects eliciting various facial expressions. These videos contain synchronized image sequences of faces in frontal and in non-frontal situations. For each subject, six head pose variations combined with seven expressions were recorded by two cameras, which results in a total of 630 constrained recordings captured with a helmet camera (i.e., without head movement) and 630 unconstrained recordings captured with a regular camera placed in front of the user (i.e., with head movements).

Concerning the above, we are using a subset of CK+ containing 374 sequences (which are commonly used in the literature). We use this subset to evaluate the 'six universal expressions recognition problem'. For SNaP-2DFe, we only use the subset acquired by the helmet camera, used to remove head pose variations. All faces from the different databases are rotated and cropped (based on 68 landmark locations), color normalized [20] and resized in order to standardize the data, as illustrated in Figure 2.

	Number of sequences	Sequences length (neutral to apex)	Number of persons	Emotion
CK+	374 **(358)**	4 to 71 img / seq **(10)**	99 **(99)**	Happiness, Sadness, Anger, Disgust, Fear, Surprise
Oulu - CASIA	480 **(478)**	9 to 72 img / seq **(10)**	80 **(80)**	
SNaP-2DFe	540 **(540)**	26 to 47 img / seq **(10)**	15 **(15)**	

Figure 2. Datasets used to analyze facial expressions from optical flow. The information in bold represents the final data obtained after the standardization process.

3.2. Temporal standardization process

As the duration of the different sequences across the datasets varies widely (from 4 to 72 images), a temporal normalization of the sequences is necessary. Two temporal normalizations have been applied: TIM2, where the optical flow is calculated only between the first image (neutral) and the last image — where the intensity of expression is at its highest (apex) — and TIM10, where 10 images are selected from the first image (neutral) to the last image (apex). For practical reasons, only sequences with at least 10 images are used in the evaluations, both for TIM2 and TIM10. In the TIM10 case, in order to select the key images within a sequence, we calculated the intra-face motion intensity induced by the expressions. To avoid considering images where head movement is more pronounced than information relating to the facial expression, we have dissociated the movement from the rigid parts of the face (contour, nose) and from the dynamic facial elements (eyebrow, eyes, mouth). The movement between two successive images can be calculated as follows:

$$f(t_1, t_2) = \begin{cases} \frac{\Delta_E + \Delta_M}{\Delta_H}, & \text{if } \Delta_H > 0. \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

where Δ_E, and Δ_M represent the motion intensity in the dynamic regions of the face (eyebrows, eyes, mouth) between the two images t_1 and t_2, while Δ_H represents the motion intensity in the rigid regions of the face (nose, contour). Following this rule, if the face is not affected by any variations ($\Delta_H = 0$) or if the head movement is too

large ($\Delta_H > \Delta_E + \Delta_M$), the value obtained will be low, implying that the associated image is not significant. An illustration of the key images selection process is shown in Figure 3. We select the n key images where the delta has changed the most between two successive images during the sequence (corresponding to the green segments in Figure 3).

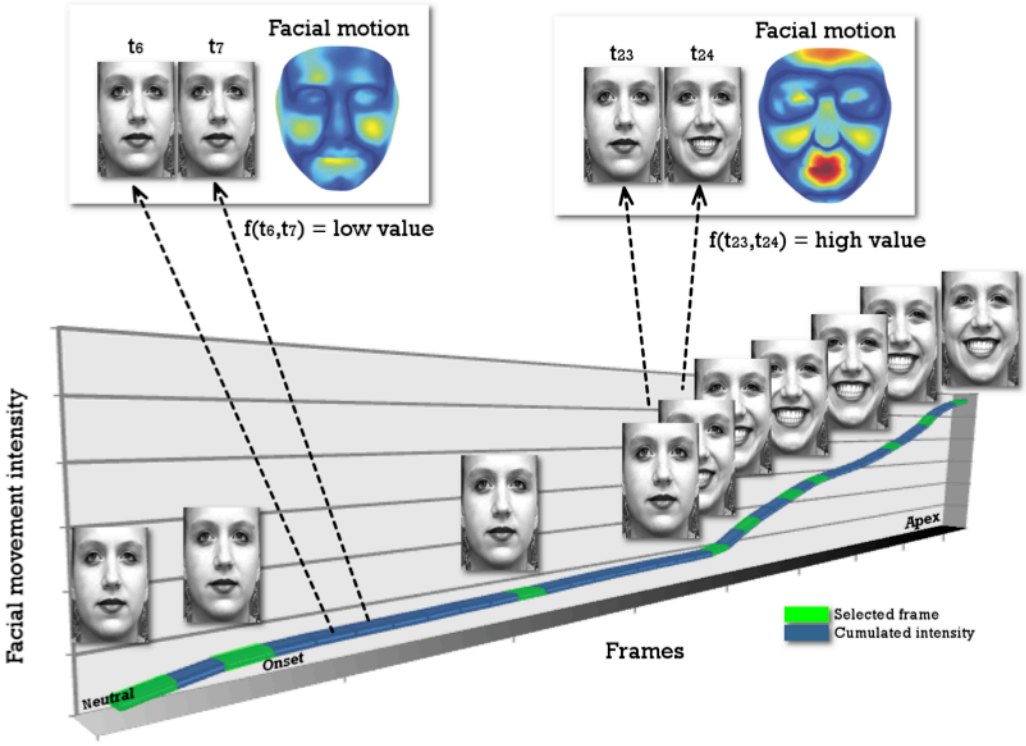

Figure 3. Selection process of the key images according to the intra-face motion.

3.3. Performance criteria

The optical flow approaches are each evaluated using SVMs of type C-SVC with linear kernels. We are aware that SVMs may not provide the highest classification accuracy, and that the reported results could be further optimized. However, our approach here is to compare optical flow approaches against a common benchmark. Hence, in the following evaluations, we focus primarily on the behavior of the different optical flow approaches and not on optimizing their performance for facial expression recognition.

To evaluate the performance of optical flow approaches for facial expression analysis, we use a 60-40 train/test validation protocol. The performance criteria being considered in our results are formulated as follows:

AUC. Suppose we have some learning examples $\{xi, yi\}_{i=1}^{l}$ and a linear decision function of the form $f(x) =< w, x > +b$. The AUC equation corresponding to this learning set and $f(x)$ is equivalent to:

$$AUC = \frac{\sum_{i=1}^{n^+} \sum_{j=1}^{n^-} f(x_i^+) f(x_j^-)}{n^+ n^-}. \quad (2)$$

where n^+ and n^- are respectively the number of positive and negative examples.

Mean AUC. In order to uniformly evaluate all optical flow approaches, we have randomly generated ten learning configurations. For each evaluation, we report the average of the AUC obtained on the different learning configurations calculated by the following equation:

$$mAUC = \frac{\sum AUC_{i=1}^{c}}{c}. \tag{3}$$

where c is the number of learning configurations (c = 10).

In the case of the *data augmentation* experiments, performance is calculated on exactly the same ten 60-40 train/test validation configurations, in order to ensure uniform evaluation of all optical flow approaches in the presence of and absence of data augmentation. For each evaluation, we select one optical flow approach and augment the training data with the remaining optical flow approaches — making sure not to take any data from the test set. We then calculate the **average accuracy** obtained for each of the configurations.

4. Evaluation of optical flow approaches

The evaluation of optical flow approaches involves the application of the standard analysis process passing through the following steps: flow estimation, expression characterization and classification. In order to avoid any bias which may suggests that one analysis system is more suitable for one optical flow approach than another, three approaches are investigated, as illustrated in Figure 4. They are as follows:

1. Analysis of the **raw** flow data input directly into the classifier;
2. Use of **hand-crafted descriptors** to build a characteristic motion vector which is then passed to a classifier;
3. Use of **deep learning architectures** which rely on learned features constructed from the available data.

4.1. Analysis of the raw flow data

In this experiment, we directly evaluate the raw flow data obtained from the different optical flow approaches. This makes it possible to verify an optical flow approach's ability to preserve facial movements without using any descriptor or encoding. For this purpose, we use a linear SVM classifier. The use of this basic classifier makes it possible to avoid more complex learning approaches that could favour a particular optical flow approach.

In this experiment, we use only the standardized sequences with TIM2 (optical flow computed between two images: neutral and apex). The values representing the characteristic vector correspond to the raw optical flow values. Each pixel is characterized by three values: two values for the direction corresponding to the cosine and sine of the angle, and one value for the magnitude of the motion. Since the images have a size of 50×50, the characteristic vector reaches a size of $50 \times 50 \times 3 = 7500$. Table 1 contains the results obtained from the various evaluations.

The results obtained by applying a basic classifier (in Table 1), suggest that there is a difference in performance between the different optical flow approaches. It is important to show that the methods which use initialization by sparse descriptor matching techniques such as Farnebäck and TV-L1 yield higher AUC values,

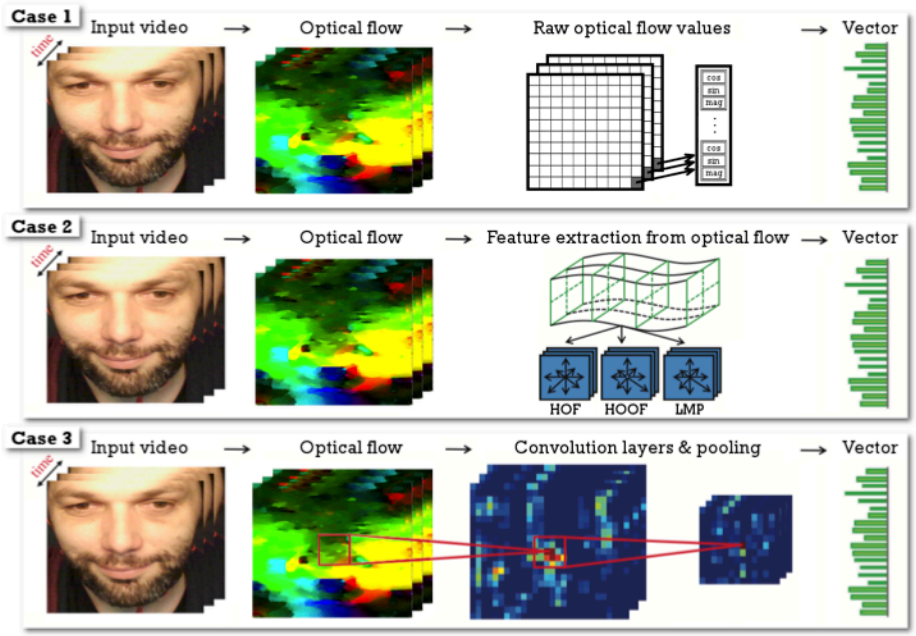

Figure 4. Comparison of analysis from raw data, hand-crafted and deep learning processes (based on optical flow and used for facial expression recognition).

Table 1. Mean AUC obtained from analysis of the raw flow data with TIM2. Performances are ranked row-wise, with the highest performances emboldened (cells are ranked visually with the darkest shades being the highest performance in the row, and the lightest shades being the lowest).

TIM2	Farnebäck	TV-L1	Ldof	Epicflow	FlowField	PWC-net
CK+	86.5% ± 2.67	**87.7% ± 2.40**	83.3% ± 3.30	54.1% ± 2.99	84.9% ± 2.07	75.6% ± 2.71
CASIA	59.2% ± 3.15	62.8% ± 2.74	56.9% ± 3.31	40.6% ± 3.47	**62.9% ± 3.03**	53.4% ± 2.31
SNaP	**75.7% ± 2.31**	67.3% ± 6.20	65.2% ± 3.01	46.9% ± 2.96	74.1% ± 2.02	66.8% 2.85

even outperforming some recent approaches, which are based on neural architectures (PWC-net). Although the FlowField approach is less effective than PWC-net on datasets such as MPI-Sintel, it stands out from other recent approaches in the context of facial expression recognition and achieves competitive results in this problem domain. This is because, unlike other recent approaches, FlowField does not require explicit regularization, smoothing (like median filtering) or a new data term. Instead it solely relies on patch matching techniques and a novel multi-scale matching strategy which appears to be better adapted for characterizing facial movement. The performance of the Epicflow approach is relatively poor in comparison, largely because this approach is not well adapted to calculate the movement between two relatively different (distant in time) images, mainly due to the matching method used.

4.2. Recognition from hand-crafted approaches

Most facial expression recognition systems use motion descriptors to more accurately characterize facial movements within the optical flow, to facilitate the classification step. To compare the performance of optical flow approaches using hand-crafted approaches, and to avoid the possible bias that some descriptors might cause on a specific optical flow approach, we use several motion descriptors that are currently used in the area of facial

expression recognition: HOF [21], HOOF [22] and LMP [1]. All these descriptors are associated with a facial segmentation model in order to characterize the global facial movement. Among the existing models, we select a classic 5×5 grid in order to avoid any bias due to an incorrect estimation of the facial regions. As a reminder, in this evaluation, we do not seek to optimize the performance of the different approaches, only to propose a fair comparison between them.

Table 2. Mean AUC obtained from the handcrafted approaches with TIM2. Performances are ranked row-wise, with the highest performances emboldened (cells are ranked visually with the darkest shades being the highest performance in the row, and the lightest shades being the lowest). Note Farnebäck's overall high performance, and Epicflow & PWC-net's overall low performance.

TIM2		Farnebäck	TV-L1	Ldof	Epicflow	FlowField	PWC-net
CK+	HOF	**77.4% ± 1.95**	76.9% ± 2.18	70.9% ± 3.44	50.1% ± 3.14	76.7% ± 2.62	66.7% ± 2.90
	HOOF	65.6% ± 2.41	62.0% ± 3.23	55.6% ± 3.16	34.6% ± 4.00	**68.5% ± 2.46**	56.2% ± 1.98
	LMP	**72.6% ± 3.09**	71.5% ± 3.77	64.5% ± 2.87	47.3% ± 1.76	64.9% ± 2.13	48.6% ± 2.67
CASIA	HOF	47.6% ± 2.17	50.0% ± 2.98	43.1% ± 3.07	32.6% ± 2.91	**53.7% ± 2.00**	41.2% ± 2.97
	HOOF	37.6% ± 2.06	39.5% ± 2.12	32.8% ± 3.55	22.9% ± 3.34	**43.2% ± 2.29**	30.5% ± 2.27
	LMP	**45.0% ± 3.29**	42.7% ± 2.26	39.0% ± 2.94	30.8% ± 3.22	39.4% ± 2.98	36.2% ± 3.88
SNaP	HOF	**63.6% ± 2.87**	48.3% ± 2.90	50.8% ± 2.69	39.0% ± 3.23	62.4% ± 1.50	55.6% ± 2.67
	HOOF	**57.3% ± 1.76**	40.4% ± 1.95	41.6% ± 3.47	31.6% ± 2.71	55.9% ± 2.80	47.0% ± 3.49
	LMP	**63.3% ± 2.75**	49.1% ± 4.43	53.4% ± 3.40	35.9% ± 2.68	59.4% ± 1.95	55.4% ± 1.71

Table 2 contains the results obtained from the evaluations of different descriptors with the TIM2 configuration (motion between the neutral and the apex image) and Table 3 with the TIM10 configuration (which takes into consideration the movement throughout the activation sequence). To account for the movement, we calculate the characteristic vector within 25 regions of the face using the descriptor. Then, we construct a temporal vector by summing the different characteristic vectors. For all the descriptors, we analyze the distribution of the local movement over 12 directions. The characteristic vector reaches a size of $12 \times 25 = 300$.

Table 3. Mean AUC obtained from the handcrafted approaches with TIM10. Performances are ranked row-wise, with the highest performances emboldened (cells are ranked visually with the darkest shades being the highest performance in the row, and the lightest shades being the lowest). Note Farnebäck's overall high performance, and PWC-net's overall low performance when compared to all other approaches.

TIM10		Farnebäck	TV-L1	Ldof	Epicflow	FlowField	PWC-net
CK+	HOF	**87.2% ± 1.87**	83.6% ± 1.57	85.2% ± 2.69	84.2% ± 3.15	86.6% ± 2.91	82.2% ± 2.61
	HOOF	**83.2% ± 2.34**	76.2% ± 2.25	78.6% ± 3.71	77.0% ± 2.30	81.4% ± 3.33	75.8% ± 1.61
	LMP	**89.7% ± 1.24**	86.8% ± 2.56	87.3% ± 3.53	86.2% ± 2.44	88.6% ± 1.48	54.1% ± 1.37
CASIA	HOF	**62.8% ± 3.11**	59.9% ± 3.60	53.8% ± 2.14	54.6% ± 2.22	62.5% ± 2.17	53.4% ± 3.59
	HOOF	**54.6% ± 1.77**	51.5% ± 2.95	45.1% ± 1.91	45.6% ± 2.59	54.5% ± 3.10	45.5% ± 3.53
	LMP	60.6% ± 2.47	65.5% ± 2.45	62.7% ± 2.43	62.4% ± 3.45	**66.1% ± 2.32**	62.5% ± 2.59
SNaP	HOF	63.9% ± 2.18	61.4% ± 1.07	56.2% ± 2.04	58.5% ± 2.75	**65.0% ± 2.90**	53.4% ± 2.41
	HOOF	**58.2% ± 2.39**	52.3% ± 3.65	49.9% ± 3.14	49.2% ± 3.08	56.2% ± 3.82	47.3% ± 0.94
	LMP	68.1% ± 2.51	68.3% ± 3.77	66.5% ± 1.77	63.6% ± 2.41	**71.1% ± 2.60**	65.0% ± 2.49

Based on the results obtained from Tables 2 and 3, two optical flow approaches repeatedly achieve the best performances: Farnebäck and FlowField. The difference in performance on the three datasets is explained by the fact that face registration is more complex on CASIA and SNAP and generates more residual noise which is reflected in the optical flow.

In Table 2, where we do not consider temporal information, FlowField and Farnebäck outperform almost all the other approaches regardless of the descriptor used, closely followed by the TV-L1 approach. As in the previous evaluation (see Table 1), the Epicflow approach gives the worst performance because it is not adapted to encode significant movements between two images.

In Table 3, it is observed that taking into account temporal information (TIM10) provides a better characterization of facial expressions. In this context, the Farnebäck approach outperforms almost all the other approaches regardless of the descriptor used. Considering the temporal information, we observe that the FlowField approach remains very competitive with the Farnebäck approach. These results show that the two approaches tend to provide more consistent movements over time than the other studied approaches. Although the performance of the Epicflow approach is always lower, it can be seen that the performance is relatively similar to the performance of the other approaches. This is because the distance between the images is less important, and the movement at the pixel level is more coherently encoded.

4.3. Recognition from using learning based approaches

In this experiment, we compare the results of different deep learning architectures when applied to different optical flows. Among the deep learning architectures used in computer vision [23], we have selected two main types of architectures: Convolutional Neural Networks (CNNs) (based on the optical flow computed from the neutral and the apex image) and Recurrent Neural Networks (RNNs) which take into account the temporal information (all images in the sequence from the neutral to the apex image). The two architectures that are used in this evaluation are shown in Figure 5. Each architecture is applied to the different datasets and optical flows.

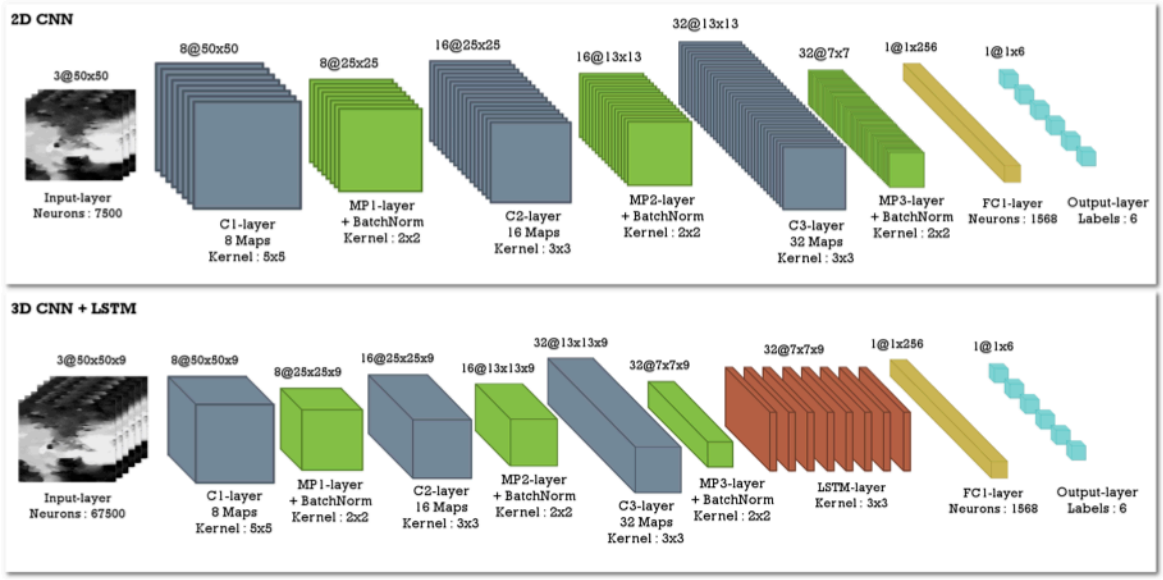

Figure 5. Neural architectures used in the evaluations (C : Convolutional layer, MP : Max pooling, FC : Fully-connected layer).

We are aware that there are other more complex architectures which produce a much better performance. However, in this evaluation, we simply wish to compare the different optical flow approaches and consider

how they perform in low complexity contexts (to minimize learning biases). For the learning data, we use the same data format as used in the evaluation in Section 4.1. Each motion pixel is characterized by three values: cosine, sine and magnitude. Since the images have a size of 50×50, the characteristic vector reaches a size of $50 \times 50 \times 3 = 7500$. For all evaluations, we use a batch size of 8 and an 10 epochs for training. Table 4 contains the results obtained for the various evaluations.

Table 4. Mean AUC obtained from the learning based approaches with TIM2 for CNN and TIM10 for RNN. Performances are ranked row-wise, with the highest performances emboldened (cells are ranked visually with the darkest shades being the highest performance in the row, and the lightest shades being the lowest).

TIM2 & 10		Farnebäck	TV-L1	Ldof	Epicflow	FlowField	PWC-net
CK+	CNN	84.7% ± 2.71	**87.7% ± 3.79**	80.2% ± 3.00	55.5% ± 3.55	84.7% ± 1.65	75.9% ± 2.86
	RNN	86.3% ± 3.11	83.7% ± 2.66	83.7% ± 2.29	82.1% ± 2.74	**87.5% ± 1.81**	81.2% ± 3.32
CASIA	CNN	55.6% ± 2.81	**62.8% ± 2.86**	53.3% ± 4.03	37.7% ± 3.13	62.6% ± 2.27	51.4% ± 2.94
	RNN	60.8% ± 2.72	62.9% ± 4.69	59.0% ± 3.52	54.7% ± 3.57	**63.1% ± 2.85**	55.5% ± 5.80
SNaP	CNN	**68.4% ± 2.41**	63.0% ± 3.12	59.9% ± 2.46	44.6% ± 2.30	68.4% ± 3.05	60.7% ± 1.90
	RNN	59.1% ± 3.55	**62.4% ± 3.16**	56.0% ± 2.16	54.0% ± 5.07	60.7% ± 3.05	54.9% ± 3.16

Considering the results obtained in Table 4, the performances of the different optical flow approaches are similar for both deep learning architectures (i.e., CNNs and RNNs). Additionally, note that the temporal information (TIM10) tends to improve the performance of approaches. Indeed, taking into account the temporal information helps in filtering out some discontinuities.

The performance of the optical flow approaches is relatively similar to the performances observed in previous evaluations. The Farnebäck, TV-L1 and FlowField methods give the best performance using both CNNs and RNNs. Once again, the Epicflow and PWC-net approaches give the worst performance.

4.4. Discussion of the optical flow evaluations

Each of these evaluations highlight the significance behind the choice of the optical flow approach in facial movement analysis — an incorrect choice can result in significantly poorer performance. To fairly compare the different optical flow approaches, all approaches have been analyzed under the same conditions, ensuring that any bias that could result from classifier optimization or model selection has been omitted.

In these evaluations, we selected different optical flow approaches which each have their own specific characteristics (See Section 2.2).

The set of results obtained on the different evaluations makes it possible to distinguish three highly performant approaches among those evaluated: Farnebäck, TV-L1 and FlowField. It is interesting to note that recent approaches such as Epicflow and PWC-net, which have proven their effectiveness on optical flow benchmarks such as MPI-Sintel, seem less efficient for facial movement analysis.

From our point of view, the major reason for the difference in performance between the analyzed optical flow approaches, is motion interpolation. Indeed, since the EpicFlow approach demonstrated its performance on the MPI-Sintel dataset, many new approaches have tended to rely on interpolation algorithms to overcome movement discontinuities. However, when analysing facial movement, some discontinuities of movement can

provide discerning information (e.g., wrinkles), which can be key in characterizing facial expressions. In this case, approaches based on a motion interpolation algorithm tend to interpret these movements as noise because there is no local consistency in the propagation of motion. Approaches such as the Farnebäck method tend to provide a more noisy optical flow but ensure that there is no bias in the estimation of facial movement. It is then more appropriate to use a filtering algorithm that allows only the coherent movement to be maintained, while avoiding discontinuities of movement.

If we were to recommend an optical flow approach that would best characterize facial movement, we would choose either the Farnebäck or the FlowField approach (see Figure 1). The main advantage of the Farnebäck approach is that it is fast to calculate, which is an important feature to have if one wants to deploy a real-time analysis system. This can be combined with a good filtering algorithm, such as the one used by the LMP descriptor [1]. This filtering algorithm is based on the properties of facial movement propagation and can be used to improve performance. As for FlowField, it is based on a rather complex matching algorithm that is relatively more computationally expensive, especially when evaluating on a CPU. Still, FlowField has shown its effectiveness on the MPI-Sintel benchmark and on characterizing facial movement.

Now, it is important to consider the relevance of calculating a perfect optical flow that would be applicable to all problems. With the large number of optical flow approaches proposed in the literature, we explore the construction of a unique augmented model which relies on a set of the most common model characteristics.

5. Data augmentation by use of optical flows

Instead of identifying the most appropriate optical flow approaches to characterize facial movement, we study whether it is possible to rely on the properties of the different optical flow approaches in order to build a unique approach for analyzing facial expressions. With the capabilities of learning-based approaches, we explore in this section whether it is possible to use different optical flow approaches to artificially augment learning data.

To assess the impact of data augmentation based on optical flow approaches, we use the CNN architecture in Figure 5 with the TIM2 configuration on the three databases which were used in the earlier experiments (see Section 4). We choose the TIM2 configuration over the TIM10 configuration, as working on sequences is much more time-consuming and memory-intensive, especially if one wants to study a multitude of data augmentation methods. Additionally, if the augmentation provides better encoding of movement information between two images, it is expected that the results should improve when considering two successive images. To ensure that the contribution of the data augmentation is accurately compared, at the expense of the performance that can be achieved, we set all random parameters consistently: the random seeds are fixed at the initialization of the learning, the initial weights of the layers and the constant biases are the same for all runs, and the learning data is fixed according to the studied configurations.

5.1. Evaluation of the data augmentation approach

In lieu of the performances obtained by the different optical flow approaches analyzed in the previous section (see Section 4), we decide to study the contribution of the data augmentation process on only two approaches: the Farnebäck and FlowField approaches. These two approaches have been selected because they tend to provide the best performance for characterizing facial movement.

For each of the datasets, we iterate through all of the different possible data augmentation configurations. The results on the CK+ dataset, CASIA dataset, and SNAP dataset are shown in Figure 6, Figure 7 and Figure 8 respectively. For each of these three figures, the first table corresponds to the results obtained by taking the Farnebäck approach as a train/test and the second table reports the results achieved by taking FlowField approach as a train/test. The different blackened boxes represent the optical flow approaches which are used for data augmentation. The first column represents the results obtained without data augmentation and the last column represents the results obtained when using a data augmentation method which uses all the studied optical flow approaches. The results for each of the configurations appear below the tables, and average accuracy is used as our performance criteria (as described in Section 3.3). The configurations shown in green represent the three augmentation configurations which obtain the highest average accuracy.

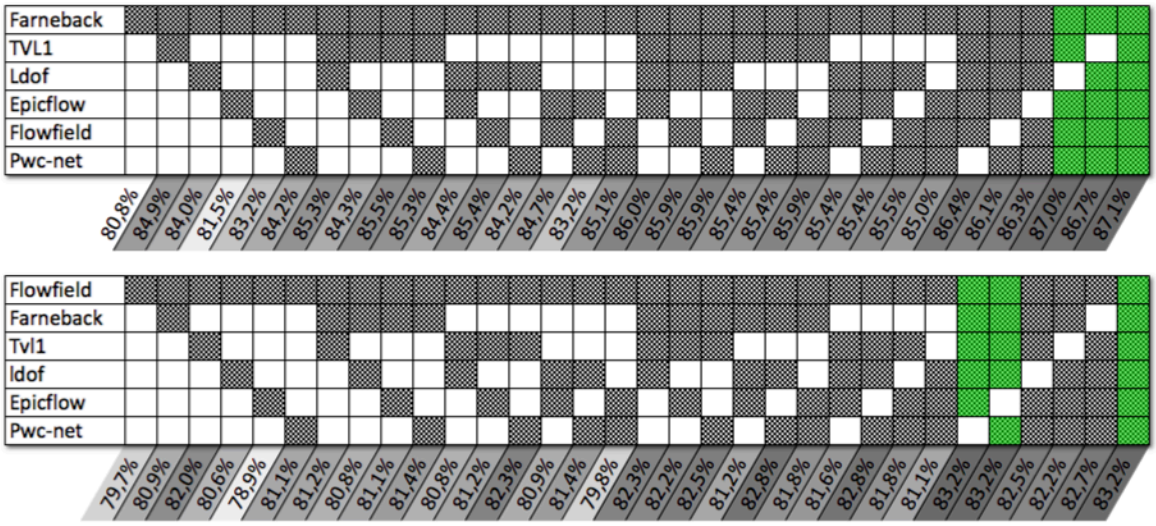

Figure 6. Data augmentation based on optical flow on the CK+ dataset (upper: Farneback; bottom: FlowField). Average accuracy is reported column-wise, for each configuration of optical flow methods used in the data augmentation process. The highest three performances are highlighted in green (figure best seen in color).

When considering the results obtained across all of the datasets, we can see that there is a significant improvement in the performance of the different optical flow approaches when the initial data are artificially augmented using other optical flow approaches. The Farneback approach gains on average 6% on CK+, 4% on CASIA and 4% on SNAP. As for the FlowField approach, it gains on average 3% on CK+, 6% on CASIA and 3% on SNAP. Overall, we notice that the more optical flow approaches that are used in the augmentation process, the more the performance tends to increase. It is thus possible to increase performance according to the

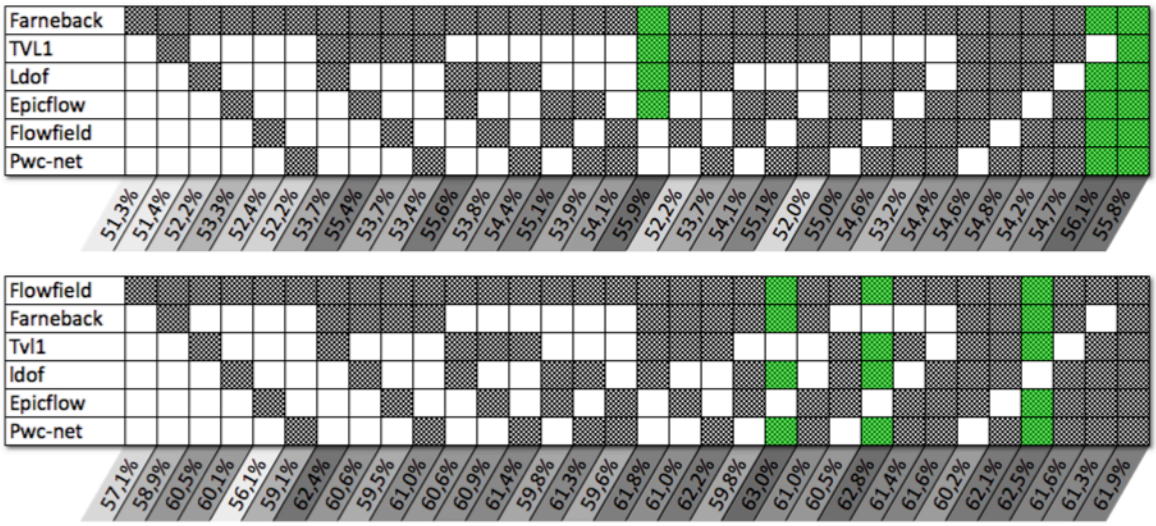

Figure 7. Data augmentation based on optical flow on the CASIA dataset (upper: Farneback; bottom: FlowField). Average accuracy is reported column-wise, for each configuration of optical flow methods used in the data augmentation process. The highest three performances are highlighted in green (figure best seen in color).

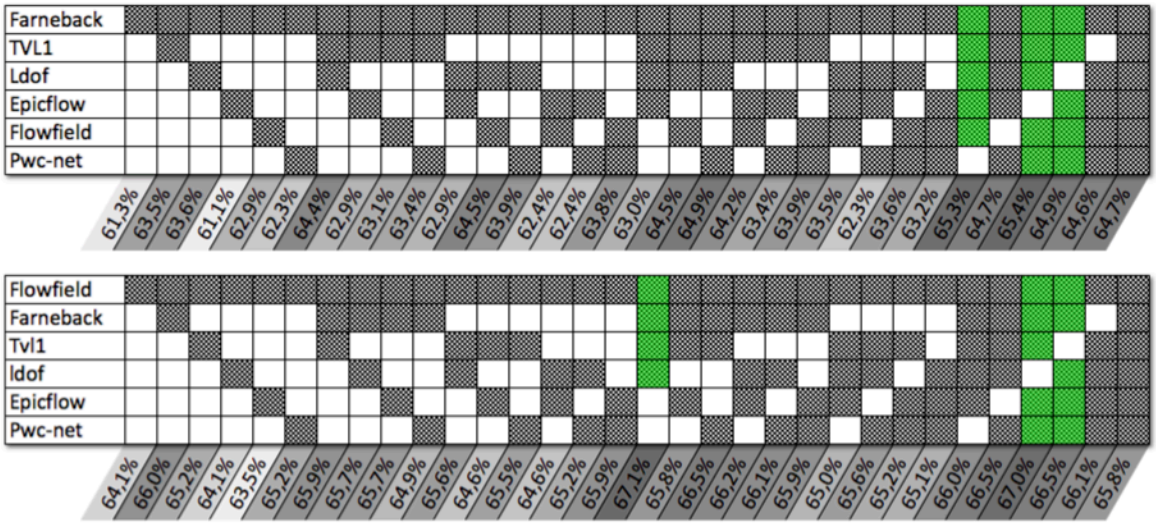

Figure 8. Data augmentation based on optical flow on the SNAP dataset (upper: Farneback; bottom: FlowField). Average accuracy is reported column-wise, for each configuration of optical flow methods used in the data augmentation process. The highest three performances are highlighted in green (figure best seen in color).

characteristics of the approaches used in the data augmentation process (robustness to motion discontinuities, illumination changes, motion intensity variations).

5.2. Discussion of the data augmentation approach

In the previous section, we investigated whether artificial data augmentation by optical flow can improve the performance of neural networks. By studying the two approaches we have identified the most suitable for analyzing facial movements (Farnebäck and FlowField), and we can see that artificial data augmentation based on other optical flow approaches can significantly improve performance (from 3% to 6% depending on the configuration).

We think it is interesting to use fast computational optical flow approaches such as the Farnebäck approach to characterize facial movement, while relying on other optical flow approaches such as FlowField to enhance learning and overcome the flaws of the less robust approaches. In the case of neural networks, it would be advisable to do offline learning with an extended set of optical flow approaches, where computation time can be relatively long. Then, use a fast but not very robust optical flow approach to extract facial movement in a real-time system.

6. Conclusion

Optical flow approaches which tend to obtain the best performance on datasets such as MPI-Sintel, seem to be less adapted to facial expression analysis than the optical flow approaches which use an initialization by sparse descriptor matching techniques, or by dense approximate fields in the nearest neighboring fields (such as the one proposed by Farnebäck). In this context, the loss of information is generally related to the way the motion discontinuities are addressed (e.g., median filtering). In this work, a primary contribution was the performance analysis of different optical flow approaches in characterizing facial expressions, and it is clear that two approaches generally outperform all the others: Farnebäck and FlowField. The Farnebäck approach has the advantage of being quick to compute, while the FlowField method has proven its effectiveness both on facial movement analysis and on more complex datasets such as MPI-Sintel.

We have thus illustrated through our experiments that some optical flow approaches differ strongly in their effectiveness in characterizing facial movements, and that it is not always easy to find a single unique solution that is both robust and fast. As such, this work's second contribution was to propose *and* benchmark a data augmentation method which uses *multiple* optical flow approaches. We have indeed shown that the artificial augmentation of a training set in this way can improve the classification accuracy. The results produced show that on average, increasing data based on optical flow approaches can improve performance by 3% to 6%, depending on the optical flow approaches used to test the data and the test dataset which is being used. This has potential applications in in-the-wild on-line analysis, where a noisy but fast optical flow can encode on the fly the data while relying on a complex offline learning process where more robust and time-consuming optical flow approaches are used for data augmentation.

In order to improve the robustness of facial optical flow, specific datasets should be released to the community. MPI-Sintel does not seem adequate for expression related challenges (variations of pose, expressions, occlusions) or challenges relating to facial motion characteristics (in the context of an expression, a discontinuity can be a source of information and does not always have to be corrected). New datasets such as SNaP-2DFE [19] which record the facial motion both in the presence or in the absence of head movements are opening the way to specific facial-expression benchmarks, but more effort should be invested in such work.

Ultimately, we believe that future work should consider the following three aspects: (1) encoding plausible facial physical constraints when extracting optical flow data, (2) the design of temporal architectures capable

of modeling the temporal activation of facial expressions and (3) exploring intra-optical and inter-optical flow augmentation techniques.

References

1. Allaert, B.; Bilasco, I.M.; Djeraba, C. Advanced local motion patterns for macro and micro facial expression recognition. *arXiv preprint arXiv:1805.01951* **2018**.
2. Yang, S.; An, L.; Lei, Y.; Li, M.; Thakoor, N.; Bhanu, B.; Liu, Y. A dense flow-based framework for real-time object registration under compound motion. *Pattern Recognition* **2017**, *63*, 279–290.
3. Poux, D.; Allaert, B.; Mennesson, J.; Ihaddadene, N.; Bilasco, L.M.; Dieraba, C. Mastering Occlusions by Using Intelligent Facial Frameworks Based on the Propagation of Movement. 2018 International Conference on Content-Based Multimedia Indexing (CBMI). IEEE, 2018, pp. 1–6.
4. Butler, D.J.; Wulff, J.; Stanley, G.B.; Black, M.J. A naturalistic open source movie for optical flow evaluation. European Conf. on Computer Vision (ECCV); A. Fitzgibbon et al. (Eds.)., Ed. Springer-Verlag, 2012, Part IV, LNCS 7577, pp. 611–625.
5. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.L. Dense trajectories and motion boundary descriptors for action recognition. *IJCV* **2013**, *103*, 60–79.
6. Snape, P.; Roussos, A.; Panagakis, Y.; Zafeiriou, S. Face flow. ICCV, 2015, pp. 2993–3001.
7. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. 13th Scandinavian Conference on Image Analysis (SCIA). Springer, 2003, pp. 363–370.
8. Cao, C.; Bradley, D.; Zhou, K.; Beeler, T. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)* **2015**, *34*, 46.
9. Wedel, A.; Pock, T.; Zach, C.; Bischof, H.; Cremers, D. An improved algorithm for tv-l 1 optical flow. In *Statistical and geometrical approaches to visual motion analysis*; Springer, 2009; pp. 23–45.
10. Kroeger, T.; Timofte, R.; Dai, D.; Van Gool, L. Fast optical flow using dense inverse search. ECCV. Springer, 2016, pp. 471–488.
11. Brox, T.; Malik, J. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence* **2011**, *33*, 500–513.
12. Revaud, J.; Weinzaepfel, P.; Harchaoui, Z.; Schmid, C. Epicflow: Edge-preserving interpolation of correspondences for optical flow. CVPR, 2015, pp. 1164–1172.
13. Weinzaepfel, P.; Revaud, J.; Harchaoui, Z.; Schmid, C. DeepFlow: Large displacement optical flow with deep matching. ICCV, 2013.
14. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)* **2013**.
15. Bailer, C.; Taetz, B.; Stricker, D. Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. ICCV, 2015, pp. 4015–4023.
16. Sun, D.; Yang, X.; Liu, M.Y.; Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8934–8943.
17. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2010, pp. 94–101.
18. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; Pietikäinen, M. Facial expression recognition from near-infrared videos. *Image and Vision Computing* **2011**, *29*, 607–619.
19. Allaert, B.; Mennesson, J.; Bilasco, I.M.; Djeraba, C. Impact of the face registration techniques on facial expressions recognition. *Signal Processing: Image Communication* **2018**, *61*, 44–53.
20. Coltuc, D.; Bolon, P.; Chassery, J.M. Exact histogram specification. *IEEE Transactions on Image Processing* **2006**, *15*, 1143–1152.
21. Essa, I.A.; Pentland, A.P. Coding, analysis, interpretation, and recognition of facial expressions. *PAMI* **1997**, *19*, 757–763.
22. Chaudhry, R.; Ravichandran, A.; Hager, G.; Vidal, R. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. CVPR. IEEE, 2009, pp. 1932–1939.
23. Khan, S.; Rahmani, H.; Shah, S.A.A.; Bennamoun, M. A guide to convolutional neural networks for computer vision. *Synthesis Lectures on Computer Vision* **2018**, *8*, 1–207.